Configurational Sampling of Noncovalent (Atmospheric) Molecular Clusters : Sulfuric Acid and Guanidine

Kubecka, Jakub

2019-07-18

Kubecka , J , Besel , V , Kurten , T , Myllys , N & Vehkamäki , H 2019 , ' Configurational Sampling of Noncovalent (Atmospheric) Molecular Clusters : Sulfuric Acid and Guanidine ' , Journal of Physical Chemistry A , vol. 123 , no. 28 , pp. 6022-6033 . https://doi.org/10.1021/acs.jpca.9b03853

http://hdl.handle.net/10138/307079
https://doi.org/10.1021/acs.jpca.9b03853

acceptedVersion

Configurational Sampling of Non-Covalent (Atmospheric) Molecular Clusters: Sulfuric Acid and Guanidine

Jakub Kubečka,[*,†,¶] Vitus Besel,[†] Theo Kurtén,[†] Nanna Myllys,[‡] and Hanna Vehkamäki[†]

†*Institute for Atmospheric and Earth System Research, University of Helsinki*

‡*Department of Chemistry, University of California, Irvine*

¶*+420 724946622*

E-mail: jakub.kubecka@helsinki.fi

Abstract

We studied the configurational sampling of non-covalently bonded molecular clusters relevant to the atmosphere. In this article, we discuss possible approaches to searching for optimal configurations, and present one alternative based on systematic configurational sampling, which seems able to overcome the typical problems associated with searching for global minima on multidimensional potential energy surfaces. Since atmospheric molecular clusters are usually held together by intermolecular bonds, we also present a cost-effective strategy for treating hydrogen bonding and proton transferring by using rigid molecules and ions in different protonation states, and illustrate its performance on clusters containing guanidine and sulfuric acid.

1 - Introduction

1.1 - Atmospheric Cluster Formation

Atmospheric aerosols are among the most interesting and challenging research topics in physics and chemistry for several reasons. They directly affect the daily lives of billions of people by degrading air quality, and thus, for instance, increasing the risk of cardiovascular and respiratory diseases.[1] Aerosol particles also play a significant role in the Earth's weather and climate, and aerosol–cloud interactions are the biggest contributor to the uncertainty in radiative forcing. Therefore, it is of utmost importance to understand the formation and growth of atmospheric aerosol particles.

A large fraction of atmospheric aerosol particles are formed via gas-to-particle conversion, in which molecular clusters bridge the gap between individual vapor molecules and newly nucleated particles. Molecular clusters have a central role in the atmospheric new-particle formation process, but exact cluster formation mechanisms remain poorly understood. The driving forces for the formation of atmospheric cluster are proton transfer reactions and hydrogen bonding interactions, whose strength determines the thermodynamic stability of the formed clusters. Gas-to-particle conversion occurs through random collisions of molecules in the gas phase. In the cluster formation process, both enthalpy and entropy are decreasing, *i.e.*, $\Delta H < 0$ and $\Delta S < 0$. Hence, although the process is thermodynamically favorable accordingly to the first law of thermodynamics (exothermic reaction), cluster formation is hindered by the second law of thermodynamics (entropy decreases).

In priciple, studying atmospheric cluster distributions involves finding all energetically low-lying structures for all relevant cluster compositions. However, a common assumption in atmospheric cluster formation studies is that the structure with the lowest Gibbs free energy (the global minimum structure) can be used for modelling, *e.g.*, cluster distributions. (See, *e.g.*, ref. 2 for an investigation of the accuracy of this assumption.) Thus, the main focus in configurational sampling of atmospheric clusters is usually the search for the global minimum.

The calculated Gibbs free energies of the global minimum structures are then further used to estimate cluster kinetics and population dynamics, for example, evaporation rates computed using detailed balance. It should be noted that evaporation rates are exponentially dependent on the Gibbs free energies, and thus even small errors the in Gibbs free energies can lead to errors of several orders of magnitude in, for instance, the modelled particle formation rates. We have previously introduced a high-level quantum chemical approach, in which geometries are optimized and vibrational frequencies are calculated using Density Functional Theory (DFT), and the electronic energy correction is calculated on top of the DFT structure using the domain-based local pair natural orbital coupled cluster method (DLPNO–CCSD(T)).[3] We have shown that this approach, here referred to as DLPNO//DFT, yields accurate Gibbs free energies even for large molecular clusters. However, another key question remains: how to find the global minimum–(free) energy configuration.

1.2 - Configurational Sampling

The vast number of possible molecular cluster configurations makes it very difficult to find the global minimum structure.[4,5] Over the last decade, searching for cluster configurations has become the main bottleneck in quantum chemical studies of atmospheric clusters with respect to both human and computer time.

Proper searching for global minimum requires exploring the multidimensional Potential Energy Surface (PES). In principle, the PES has $3n-6$ coordinates, where n is the number of atoms in the cluster. Moreover, at non-zero temperature, the entropy effect has to be included in the calculations as well. Unfortunately, it is not possible to use simple "brute-force" approaches such as sampling on a $(3n-6)$-dimensional grid (too large number of combinations even for modest values of n), sampling from stochastic or simple brute-force Monte Carlo (MC) simulations (sticking within one or few minima for too long time) *etc.*; as all of them are simply computationally too expensive. Various promising techniques have therefore been developed for exploring the PES at lower cost but with a sufficient

accuracy: basin hopping,[6] umbrella sampling,[7] neural networks,[8,9] or genetic algorithms (GA) methods, which have recently been shown to be quite successful.[10–14]

Furthermore, exploration of a PES using just high-level quantum chemical methods is computationally very expensive. Therefore, configurational exploration at low level of theory is utilized. We present a systematic approach for configurational sampling based on a "building up" approach.[15] This approach is commonly used in configurational sampling of proteins from single amino-acids.[16] Several structures of local minima of clusters (or proteins) are found on a PES described by methods of Molecular Mechanics (MM). MM uses classical mechanics, Force Field (FF), to describe molecular interactions. Such structures are already geometrically close to the real structures. However, they still have to be optimized on a higher level of theory. First, some redundant candidate structures can be eliminated by optimization using a low level of theory (*e.g.*, semi-empirical or tight-binding DFT methods). Thus, the number of optimizations and/or energy evaluations that need to be performed at a high level of theory (typically DFT or some wave-function-based method) can be reduced by performing a sequential series of calculations, and filtering out most structures already during the computationally cheaper stages. Since configurational sampling of molecular clusters is mainly about forming clusters with different binding patterns from a relatively small set of different monomers (single molecules), the same approach as for proteins can be used here. Figure 1 illustrates three schematic PESs of the same system at different levels of theory. The positions and depth of wells for the lowest theory (bottom PES) differs slightly from other methods, but, they still have many similarities. After first optimization (middle PES), the structures already have quite good geometries, however, the energy description has to be provided by some high-level quantum chemical method (top PES).

As mentioned above, genetic algorithm (GA) methods seem to be promising tools for exploring multidimensional PES. GA methods have been implemented in the programs OGOLOM[17] and CLUSTER,[18] and later applied for the configurational sampling of molecular clusters by Temelso *et al.*[19] and Kildgaard *et al.*[20] Karaboga[21] proposed the Artificial

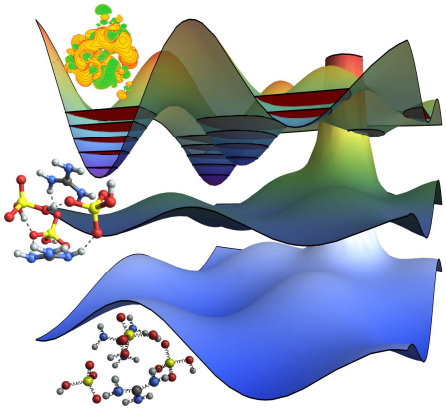


Figure 1: Illustrative scheme of mapping Potential Energy Surfaces (PES) using quantum chemical (top), semi-empirical (middle) and force-field (bottom) methods.

Bee Colony (ABC) algorithm for exploring multidimensional spaces, which was later implemented in the ABCluster program,[22,23] with focus on configurational sampling of systems composed of multiple atoms or molecules. The ABC algorithm has been applied to molecular clusters containing molecules such as ammonia,[24] ammonia and nitric acid,[25] and a series of amides and sulfuric acid.[26]

In this paper, we discuss some critical issues which have to be taken into account to ensure proper configurational sampling. We also present a protocol which we have found to be sufficient for configurational sampling of hydrogen-bonded molecular clusters containing up to 8 molecules. We selected the sulfuric acid–guanidine system to demonstrate our configurational sampling process. This system is rich in hydrogen bond donor and acceptor groups, leading to a very large number of different possible bonding patterns even for a modest number of molecules, which makes configurational sampling challenging.

This work is divided into three subsections. First, we present computational methods involved in the configurational sampling protocol itself, and a detailed description and analysis of each optimization step. Second, we demonstrate its application on the sulfuric acid–guanidine system and compare our results with results presented in previous publication.[3] Finally, we discuss the effect of proper configurational sampling on both the sulfuric acid–guanidine system, as well as other molecular clusters.

2 - Computational Methods

2.1 - Configurational Exploration

2.1.1 - Molecular Mechanics Methods

Atmospheric molecular clusters are mostly held together through Coulomb interactions and hydrogen bonds. Since structures of the individual molecules inside of clusters do not change significantly, the molecules can be treated as rigid bodies in the initial step of exploring the Potential Energy Surface (PES). Moreover, the rigid molecule approximation also avoids the possibility of undesired chemical reactions occurring during the PES exploration. Coulomb interactions are the most significant factor for self-organizing polar molecules in clusters. Thus, a configurational sampling algorithm based on Force Field (FF) methods describing these interactions in terms of partial atom-center charges can be utilized. Such approaches are indeed part of many existing configurational sampling programs.[17,18,22,23]

In our approach, the orientation and position of all rigid molecules are optimized to minimize the total intermolecular energy. The intermolecular interaction terms of force fields practically always contain a repulsive part, which prevents overlapping of atoms or molecules. A typical example is the Lennard-Jones potential E_{ij}^{LJ} which is given for atoms i and j with distance r_{ij} by

$$E_{ij}^{\mathrm{LJ}}(r_{ij}) = \epsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right), \tag{1}$$

where ϵ is the energy well depth and σ is the lowest non-overlapping distance of interacting atoms. (Here, the r_{ij}^{-12} term corresponds to the repulsive part, and the r_{ij}^{-6} term corresponds to attractive, but rather weak dispersion interactions.) For polar molecules, the formation of different bonding patterns are mainly driven by Coulomb interactions E_{ij}^{C}

$$E_{ij}^{\mathrm{C}}(r_{ij}) = k_{\mathrm{e}}\frac{q_i q_j}{r_{ij}}, \tag{2}$$

where $k_e \approx 9 \cdot 10^9\,\mathrm{Nm^2C^{-2}}$ and q represents a (typically atom-centered) partial charge.

In this work, we use the CHARMM (Chemistry at Harvard MM) FF parameters.[27,28] Structures and parameters for those molecules which are not included in the CHARMM parameters can be obtained by optimizing the structures at the MP2/6-31++G(d,p)[29–33] level of theory. Partial charges can then be extracted using the Natural Bond Orbital (NBO)[34] population analysis at this level. Missing parameters, such as the Lennard-Jones energy ϵ or distance σ, were taken from similar structures presented in the CHARMM force field database.[27,28] This approach is generally sufficient for configurational sampling of atmospheric molecular clusters. When necessary, system-specific FF molecular parameters can also be developed more rigorously, as described in ref. 35 and 36.

2.1.2 - Protonation State/Conformer Prediction Algorithm

The rigid molecule approximation made above prevents proton transfer between acids and bases. Consequently, some low-energy conformations might not be found at all. Therefore, all possible protonation states, as well as different conformational structures of the studied molecules, have to be introduced as different input structures. This often leads to a certain degree of redundancy, as the same hydrogen bond can be described by both A-H$\cdots$B and A$\cdots$H-B, depending on whether the simulation was run with either AH and B, or A and HB, as rigid bodies. The correct structures (hydrogen bond lengths and positions) are reached after optimizations with methods describing proton transfer (*e.g.*, semi-empirical or quantum chemistry methods).

Figure 2 illustrates how different protonation states/conformers could be introduced for the configurational sampling of a cluster containing 1 sulfuric acid (sa) and 1 guanidine (gd) molecule. A guanidine molecule has two protonation states: neutral (guanidine, gd = CN_3H_5) and its protonated form (guanidinium, $gd^+ = CN_3H_6^+$). In the case of sulfuric acid, not only different protonation states but also different conformers have to be introduced (hydrogen sulfate, $sa^- = HSO_4^-$; "*trans*"- and "*cis*"-sulfuric acid, sa = H_2SO_4). Thus, to

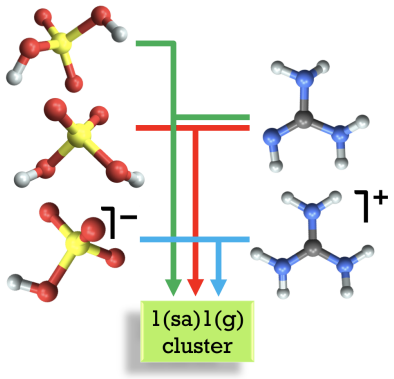


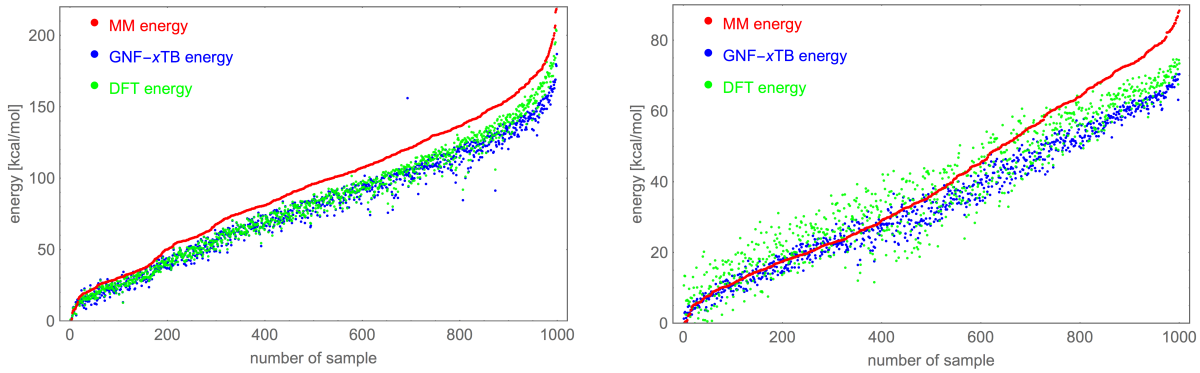
Figure 2: An illustrative scheme of different protonation states/conformers participating in the construction of 1(sa)1(gd) clusters. Atom representation: S = yellow, O = red, N = blue, C = grey, H = white.

construct a molecular cluster, all possible combinations of these molecular states have to be taken into account (while keeping the overall charge of the cluster constant). With an increase of cluster size (number of molecules), the number of possible combinations also increases. However, even the relatively large cluster 4(sa)4(gd) requires only 27 different combinations.

2.1.3 - Force-Field Based PES Exploration

In order to test the performance of Molecular Mechanics (MM) for atmospheric molecular clusters, we performed configurational sampling of the 2(sa)2(gd) cluster with two different approaches, as described in section **2.1.2 - Protonation State/Conformer Prediction Algorithm**, using either 2×("*trans*"-sa)+2×(gd) or 2×(sa^-)+2×(gd^+) as the rigid bodies. 1000 randomly generated and optimized structures were saved from the ABCluster program in both cases. As an illustration of the MM energy performance, single point electronic energy calculations were performed with both a semi-empirical method (GFN-xTB[37]) and with a quantum chemistry method (DFT: ωB97X-D[38] with the 6-31++G(d,p) basis set). Figure 3 shows the correlation between all three methods. When the system is composed of ionic rigid bodies, the MM description provides satisfying correlation with high levels of theory (see figure 3a). The partial charges of the force-field method are able to represent the

strong binding between anions and cations relatively well. The force-field description does not contain terms for, *e.g.*, the polarization of neutral molecules, and thus fails to accurately describe the strong hydrogen bonds between neutral acid and base molecules. Thus, in the case of neutral rigid bodies, the correlation of the three methods (see figure 3b) is not as strong as in the previous case (figure 3a). However, we can still conclude that structures with low energy at high levels of theory also tend to have low energies at the MM level.



(a) Cluster consisting of 2 guanidinium (gd^+) and 2 hydrogen sulfates (sa^-) ions.

(b) Cluster consisting of 2 guanidine (gd) and 2 sulfuric acid (sa) molecules.

Figure 3: Relative single point energies of 1000 randomly generated and optimized structures by the ABCluster program in two different protonation cases. Energies are related to global minima in each method. All samples are sorted according to increasing MM energy.

Based on results described in the previous paragraph, we do not have to save all minima found by the MM configurational search: energetically high-lying structures may safely be filtered out. To illustrate the usefulness of this filtering, we performed exhaustive configurational sampling of molecular clusters containing up to 4 molecules (all possible combinations of sulfuric acid and guanidine clusters) followed by GFN-xTB optimization. Figure 4 illustrates the amount of local minima which need to be optimized in order to find a certain percentage of all possible local minima of a particular molecular cluster. All minima of small clusters can be found with little effort. However, when the cluster size increases, the complete exploration of the PES becomes impossible. Fortunately, since MM energies correlate with energies calculated by quantum chemistry methods, the amount of structures required for further analysis is significantly reduced. Moreover, studying all minima would not be

even possible for larger systems, since the amount of local minima scales with cluster size N at least as $\mathcal{O}(e^N)$.

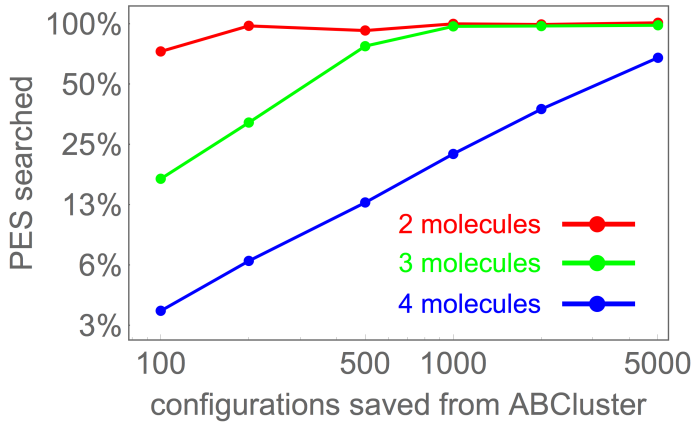


Figure 4: Illustrative picture of the number of structures which must be extracted from configurational sampling (by the ABCluster program) to find a certain percentage of all possible local minima on the GFN-xTB potential energy surface (note the logarithmic axes).

2.2 - Towards Quantum Chemistry Calculations

2.2.1 - Pre-optimization Step: Low Level of Theory

Molecular Mechanics (MM) configurational sampling of rigid molecules provides structures which require further optimization. At a higher (*e.g.*, semi-empirical or DFT) level of theory, the molecules in the MM-generated structures are under tension, because of their previously assumed rigidity, and because the hydrogen bonds between molecules are not relaxed to their minimum-energy positions due to the limited ability of the FF to describe H-bonding. Configurational sampling of clusters, where H-bonds and proton transfers are present and the configurations are mainly driven by them (typical for atmospheric clusters), require correct identification of H/proton positions. Thus, carefully selecting an appropriate approach for dealing with H-bonding and proton transfer is very important. Performing quantum chemistry optimization immediately would require many optimization steps (each associated with several self-consistent field steps to find the energy) to fully relax the whole molecular cluster. Moreover, there are hundreds to thousands of configurations provided by MM configurational

sampling, and some of them are redundant as they may end up in the same local minimum after further optimization. Therefore, we strongly recommend pre-optimization using computationally fast semi-empirical chemistry methods such as GFN-xTB method[37] (or the new GFN2-xTB version[39]), PM6,[40] or PM7.[41] Semi-empirically optimized structures also represent local minima quite similar to the "real" (*e.g.*, DFT) minima. The list of structures taken to further quantum chemistry calculations might then also be reduced by filtering redundant structures due to similarity (see section **2.2.2 - Uniqueness, Filtering and Sampling**).

It should be noted that semi-empirical methods might not be well parameterized for all molecular systems (*e.g.*, some radical systems or reactive systems). Thus, optimization of these system at a semi-empirical level could lead to non-physical structures, or even to the formation of impossible (or at least unwanted) new bonds. For these systems we recommend skipping this step, and focusing on a narrow selection of a representative set of structures already from the MM configurational sampling (see section **2.2.2 - Uniqueness, Filtering and Sampling**). One could also perform just single point calculations with semi-empirical methods to obtain inter-comparable energies (*e.g.*, for filtering) also for clusters with different sets of rigid-molecule building blocks, however, we suggest to use other collective coordinates since the energy evaluated at a different level than the level at which the structure was optimized might lead to erroneous assumptions or results. This energy calibration is needed as comparing total intermolecular MM energies from configurational sampling for clusters consisting of different building blocks (= protonation states/conformers) does not make sense as their reference energies are different.

2.2.2 - Uniqueness, Filtering and Sampling

In each step of a configurational sampling protocol, we can exclude redundant structures, and thus save computation time. One possible strategy for identifying redundant structures is to calculate the Mean-Squared-Deviation (MSD, or Root-MSD (RMSD))[42] of two different

configurations A and B (centered and oriented in the same manner[43,44])

$$\mathrm{MSD}_{\mathrm{A,B}} = \frac{1}{N}\sum_{i=1}^{N} |\vec{r}_{i,\mathrm{A}} - \vec{r}_{i,\mathrm{B}}|^2, \tag{3}$$

where $\vec{r}_{i,\mathrm{X}}$ is the position of atom i in configuration X, which has overall N atoms. Performing pair-wise comparisons of MSD for more than thousands of configurations becomes computationally exhausting especially due to non-trivial centering and orienting two clusters in the same manner. Therefore, we recommend comparing only those structures which have similar values for some suitable (easily computed) collective coordinate(s).

Collective coordinates quantify some properties of molecular clusters, and can thus help to distinguish between two different structures. For describing cluster configurations, we use a collective variable called the radius of gyration R_g (inspired by polymer science[45])

$$R_\mathrm{g}^2 = \frac{\sum_{i=1}^{N} m_i |\vec{r}_i - \vec{r}_\mathrm{COM}|^2}{\sum_{i=1}^{N} m_i}, \tag{4}$$

where m_i is the mass of atom i, $\vec{r}_i$ is its position and $\vec{r}_\mathrm{COM}$ represents the centre of mass of the whole cluster. However, it is possible to use other collective coordinates such as electronic energy, dipole moment or amount of hydrogen bonds (see Principal Component Analysis (PCA)[46,47]).

UNIQUENESS: Two molecules or clusters can be assumed to be the same if all selected collective coordinates differ by less than some threshold. We use thresholds of 0.01 Ångström, 0.001 Hartree and 0.1 Debye for the radius of gyration, the energy and the dipole moment, respectively. Smaller systems or more detailed global minimum searches may require lower energy thresholds.

FILTERING: Structures which have high energies with respect to the global minimum can also be omitted from higher-level calculations. The left-hand side of figure 5 illustrates unique (defined as above) minima of 2(sa)2(gd) clusters plotted using two collective coordinates (radius of gyration and relative electronic energy) after semi-empirical optimization. If we

assume that the semi-empirical structure corresponding to the global minimum at the DFT level is within, for example, 30 kcal/mol (illustrative arbitrary number) of the energetically lowest-lying semi-empirical configuration, we can remove all structures above 30 kcal/mol (see blue dashed line in the graph).

SAMPLING: After processing uniqueness and filtering, a large amount of structures might still remain. One can uniformly select (sample) just a representative amount of points covering the surface described by all (2 or more) collective coordinates.[48] Figure 5 illustrates sampling using two collective coordinates for configurations of a 2(sa)2(gd) cluster. The energetically lowest-lying structure is selected for each combination of collective coordinates (red point) and all neighbouring structures are removed (gray circle/ellipse). This step is repeated until no more points can be selected. Figure 5 shows a selection of 25 points. The amount of points can be varied by elongating the ellipse axis.

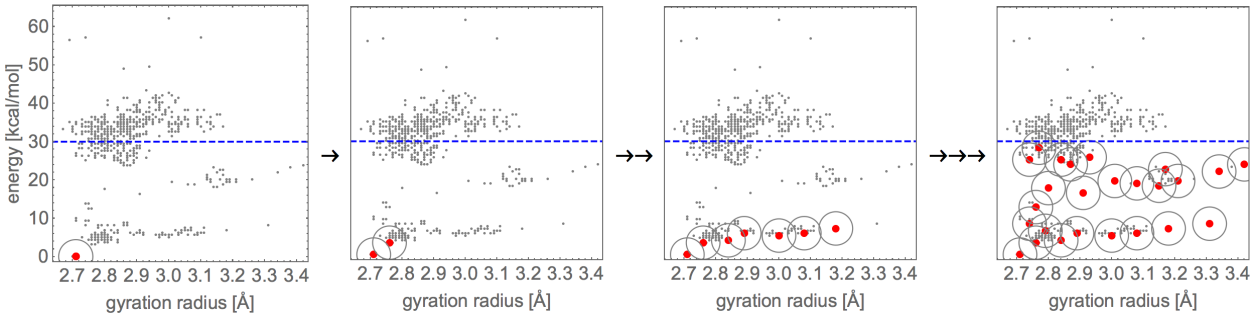


Figure 5: Scheme for sampling a representative set of structures from a PES for further analysis. The points represent different geometry minima of a 2(sa)2(gd) cluster. The energy and radius of gyration of structures are evaluated after optimization at the GFN-xTB level.[37]

2.3 - Calculation of Molecular/Cluster Properties

2.3.1 - Final Optimization, Electronic Energy Corrections, and Thermodynamic Properties

The selected structures have to be optimized at a quantum chemistry level. Note that different quantum chemical methods might lead to different global minima. Even if the bonding patterns of the global minima are the same, the bond lengths vary, and re-optimization at the

desired level has to be performed (for example, when structures are extracted from previous studies).

Density Functional Theory (DFT) with functionals M06-2X,[49] PW91,[50] ωB97X-D,[38] or PW6B95-D3[51] has been successfully applied to atmospheric molecular clusters in recent years.[20,52–54] In this work we use just the 6-31++G(d,p) basis set because DFT calculations are followed by electronic energy corrections. This basis set has been tested to be sufficient for geometries and vibrational frequencies of molecular clusters.[52] However, if DFT is used to calculate final Gibbs free energies, a larger basis set is needed to reach converged binding energies.

For configurational sampling, we suggest that an initial DFT optimization is performed with loose convergence criteria, followed by sequential optimizations with tighter convergence criteria (and where needed, *e.g.*, finer integration grids) to obtain the final DFT electronic energy $E_{\mathrm{el}}^{\mathrm{DFT}}$. Again, between these optimization steps, a new filtering (and if needed also sampling) can be processed to remove redundant structures. For the optimized structures, a vibrational frequency analysis should then be performed to obtain the Gibbs free energy G^{DFT} at a desired temperature. Correctly optimized structures do not contain any imaginary frequencies, but might contain low-lying frequencies (lower than 100-200 cm^{-1}). These "vibrational frequencies" may originate from internal rotations of molecules or functional groups within the cluster, and should not be treated as vibrations in the calculation of partition functions. One can use the quasi-harmonic approximation to recalculate frequencies (implemented in the program GoodVibes[55]) or just simply replace low-lying frequencies (<X cm^{-1}) by some cut-off value X cm^{-1}.[56,57] The anharmonic corrections would slightly decrease free energy of cluster, however, the correction is much smaller in amplitude compared to the quasi-harmonic approximation.[52] Due to its low magnitude, the anharmonic corrections do not affect the configurational sampling, and thus we do not apply them in this article. If necessary, a single point energy calculation at, *e.g.*, Coupled-Cluster (CC) level of theory can also be performed to obtain a correction to the electronic energy $E_{\mathrm{corr}}^{\mathrm{CC}}$. The corrected

Gibbs free energy then has the form

$$G = G^{\mathrm{DFT}} - E_{\mathrm{el}}^{\mathrm{DFT}} + E_{\mathrm{corr}}^{\mathrm{CC}}. \tag{5}$$

Molecular clusters have many relevant low-lying configurations, and when necessary the Gibbs free energy can also be computed for an ensemble of conformers as[2]

$$G = -RT\ln\Big(\sum_i e^{-G_i/RT}\Big), \tag{6}$$

where R is the universal gas constant, T is the temperature and G_i represents the Gibbs free energy of conformer i. Partanen *et al.*[2] showed that the effect of multi-configurational averaging is modest compared to, *e.g.*, error sources in $E_{\mathrm{el}}^{\mathrm{DFT}}$, and thus the global minima search remains the main focus of our configurational sampling. We confirm Partanen's conclusion for the system of sulfuric acid–guanidine in Supporting Information.

Finally, the formation (binding) Gibbs free energy ΔG of a cluster can be calculated as

$$\Delta G = G_{\mathrm{cluster}} - \sum_i G_{\mathrm{monomer}\,i}\,. \tag{7}$$

We would like to also point out that some quantum chemistry programs do not correctly classify the symmetry point groups of molecules/clusters due to a low numerical threshold in the symmetry identification. For instance, molecules like water, ammonia, guanidine or sulfuric acid should have the assigned symmetry $C_{2v}(\sigma_{\mathrm{R}} = 2)$, $C_{3v}(\sigma_{\mathrm{R}} = 3)$, $C_1(\sigma_{\mathrm{R}} = 1)$ and $C_2(\sigma_{\mathrm{R}} = 2)$ respectively, where σ_{R} represents the rotational symmetry number.[58,59] External programs can be used to check the symmetry of a molecule/cluster within adjustable threshold.[60] Where necessary, corrections for the incorrect rotational symmetry number σ_{R} should then be performed, as otherwise the partition function double counts several micro-

states of the same type. The corrected expression is then

$$G = G^{C1} + RT\ln(\sigma_{\mathrm{R}}), \tag{8}$$

where G^{C1} represent the Gibbs free energy calculated with no symmetry. Consequently,

$$\Delta G = \Delta G^{C1} + RT\ln(\frac{\sigma_{\mathrm{R,cluster}}}{\prod_i \sigma^i_{\mathrm{R,monomer}\,i}}), \tag{9}$$

where $\sigma_{\mathrm{R,cluster}}$ corresponds to the point group of the cluster, as also clusters may be symmetric.

In this article, we have assumed that the global minima of clusters of sulfuric acid and guanidine behave as crystals due to the dense network of strong hydrogen bonds. In other words, we assume that molecules do not easily exchange positions with each other within the cluster. If molecules inside a cluster are easily interchanged, the total symmetry number, which is much more complicated to calculate (the indistinguishability of molecules have to be assumed), would have to be used instead of the rotational symmetry number.[58,61,62] However, we believe that this assumption is correct at least for the most strongly bound clusters with equal number of sulfuric acid and guanidine molecules, which are the most important structures for the studies of new-particle formation in the atmosphere.

3 - Results

In this section, we first present all global minima found in our research, and compare them with global minima found in previous study using a different approach. Next, we present a universal protocol for configurational sampling of hydrogen-bonded molecular clusters, and describe the details of all steps involved in it. We also illustrate the performance of our protocol. Finally, we briefly discuss symmetries of global minima clusters and the effect of symmetry on the formation Gibbs free energy.

3.1 - Global Minima

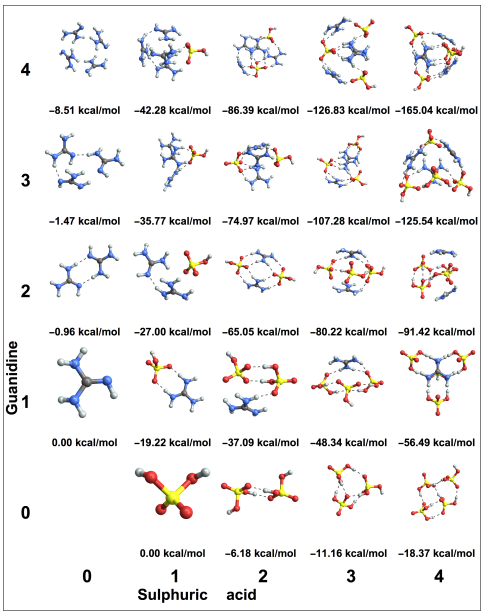


Figure 6: Geometric structures of global minima for clusters containing a mixture of guanidine and sulfuric acid molecules. The global minima are found with respect to DLPNO//ωB97X-D/6-31++G(d,p) PES. The Gibbs free formation energies are shown below each structure. Atom representation: S = yellow, O = red, N = blue, C = grey, H = white.

We have developed a program named *Jammy Key for Configurational Sampling* (JKCS), which is composed of sets of scripts operating with a large amount of files (structures *etc.*), communicating with different quantum chemistry computational programs, and managing all required calculations on a supercomputer via the SLURM scheduler. The JKCS program is not available online yet, but can be obtained by contacting the authors.

We utilized JKCS to find global minima of sulfuric acid–guanidine molecular clusters at room temperature T = 298.15 K. The free energy of the clusters is computed using a range-separated hybrid density functional with an empirical dispersion correction, ωB97X-D,[38] and the 6-31++G(d,p) basis set. Low frequencies were treated with the quasi-harmonic approximation using a frequency threshold of 100 cm^{-1}. On top of the DFT optimized structures, we corrected the electronic energy using the Domain-based Local Pair Natural Orbital Coupled Cluster (DLPNO–CCSD(T)) method[63–66] with an aug-cc-pVTZ basis set.[67,68] The

linear scaling DLPNO–CCSD(T) method and the Tight Pair Natural Orbital (TightPNO) criteria are used, as recommended for non-covalently bound systems.[69] Henceforth, we refer to this combination using the shorthand notation DLPNO//ωB97X-D/6-31++G(d,p).

Configurational sampling was performed using the ABCluster program.[22,23] Intermediate re-optimizations were done using GFN-xTB.[37] All DFT calculations (structure optimization and vibrational frequency analysis) were performed with Gaussian 16 Revision A.03.[70] Low vibrational frequencies were corrected via the GoodVibes program.[55] DLPNO calculations were performed with the Orca program version 4.0.1.2.[71]

We have tested various sequences of optimization steps, filtering steps, sampling/selecting procedures *etc.* As result of this, we have found a large number of local minimum structures. The lowest minima found are assumed to be the global minima and are presented in figure 6. The XYZ coordinates and quantum chemistry programs outputs are also given in the Supporting Information. In figure 6, the formation energies of clusters (excluding the monomer of sulfuric acid, σ(sa)= 2) are computed with the assumption that all clusters have a symmetry number of 1 (see section **3.3 - Symmetry Contribution** for further discussion of symmetry).

To compare our results with previous study, we compare the Gibbs free energies of global minima presented by Myllys *et al.*[3] who have studied the sulfuric acid–guanidine system using a configuration sampling approach presented by Elm *et al.*[72] Figure 7 shows the difference of Gibbs free energies between our and their results. In most cases, we have found the same global minimum. However, we can show that several cases, structures with energies lower than 0.5 kcal/mol compared to previous studies have been found (the greatest improvement reaches almost 8 kcal/mol).

The structure of $(sa)_1(gd)_3$ represents an exceptional case. The global minimum shown in figure 8a has been taken from a previous study by Myllys *et al.*[3] because the lowest structure that we have found (see figure 8b) is 0.05 kcal/mol higher in free energy than the structure presented by Myllys *et al.* However, the structures are almost identical with very small dif-

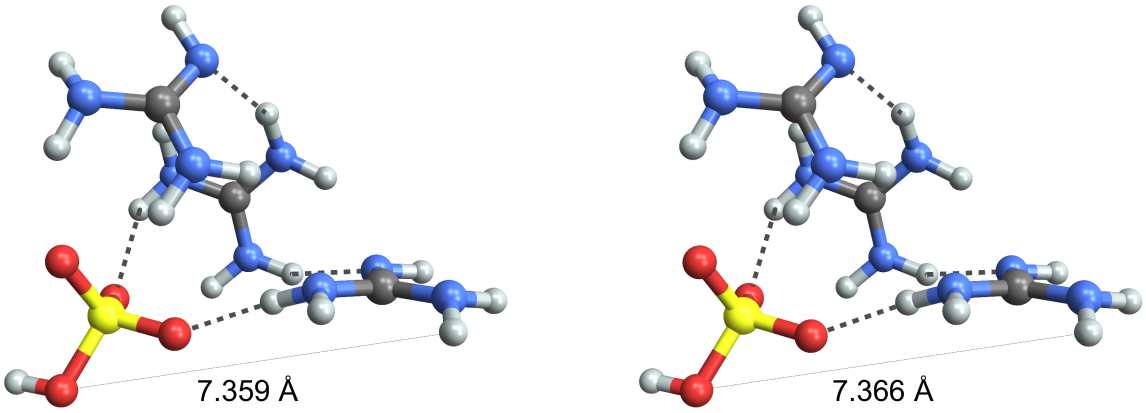
GUANIDINE	4	-0.05	-0.22	-7.74	0.00	-3.87
	3	0.00	0.05	-4.47	-0.69	0.00
	2	0.00	0.00	0.00	-0.19	-1.92
	1	0.00	-0.13	-0.02	-0.01	-0.54
	0		0.00	0.00	-0.99	-2.03
		0	1	2	3	4
in [kcal/mol]		SULPHURIC ACID				

Figure 7: Difference of DLPNO//ωB97X-D/6-31++G(d,p) Gibbs free energies (at 298.15 K) between global minima found by our studies and structures shown as global minima in previous study.[3] Negative value means that we have found better structure with energy lower than X kcal/mol, and *vice versa*. The green color background highlights improvements greater than 0.5 kcal/mol. The red color highlights cases where we were unable to find the global minimum reported in previous studies.

ferences in bond lengths and molecule orientations. The reason that our approach does not find this minimum is likely that the corresponding structure is removed during the uniqueness check stages, because both structures are the same minimum based on our uniqueness evaluation parameters (see section **2.2.2 - Uniqueness, Filtering and Sampling**). Moreover, the energy difference of 0.05 kcal/mol is small enough that the lowest-energy structure found by our approach can be considered a good approximation of the global minimum - the error from the configurational sampling is at least an order of magnitude smaller than the errors associated with the quantum chemical methods (including especially the calculation of entropies).

3.2 - Universal Protocol for Configurational Sampling

Using the set of global minima for sulfuric acid–guanidine described above, we next strive to develop a universal protocol which can find these minima as cost-effectively as possible, and which can easily be adapted to clusters containing other molecules. Table 1 lists all the individual steps of our protocol, which are described in detail in the following paragraphs.



(a) Global minimum of $(sa)_1(gd)_3$ found by Myllys *et al.*[3]

(b) The lowest minimum of $(sa)_1(gd)_3$ found in this work.

Figure 8: Comparison of the two "different" minima structures $(sa)_1(gd)_3$. The lowest minimum (figure 8b) found in this work is over 0.05 kcal/mol higher than the global minimum (figure 8a). Atom representation: S = yellow, O = red, N = blue, C = grey, H = white.

To illustrate the typical computational cost of different jobs or different configurational sampling steps, the computational times (= computational cost in cpu-hours) from table 1 are visualized in figure 9.

Table 1: The universal protocol for configurational sampling. The variable N corresponds to the number of molecules in the molecular cluster.

METHOD	Amount of INPUT structures	JOB TIME $[\frac{\text{hours}}{\text{job}\times\text{CPU}}]$	TOTAL TIME $[\frac{\text{hours}}{\text{CPU}}]$	ENERGY FILTER THRESH. [kcal/mol]	Amount of OUTPUT structures selected
search by ABCluster	-	-	$\approx 1N^4$	-	$\approx 2.5N10^4$
GFN-xTB pre-optim.	$\approx 2.5N10^4$	$\approx 0.83 \cdot 10^{-4}N^2$	$\approx 2N^3$	$<5\,N$	$\approx 100N$
low DFT (loose opt.)	$\approx 100N$	$\approx 22 \cdot 10^{-4}N^3$	$\approx 0.22N^4$	$<2.5\,N$	all
high DFT (v.tight opt.)	$\approx 20N$	$\approx 125 \cdot 10^{-4}N^3$	$\approx 0.25N^4$	$<1.7\,N$	all
high DFT (freq. calc.)	$\approx 15N$	$\approx 27 \cdot 10^{-4}N^3$	$\approx 0.04N^4$	<1	≈ 3
DLPNO (SP calc.)	≈ 3	$\approx 2.4N^3$	$\approx 7.2N^3$	-	-

First, we perform a PES exploration at the MM level using the ABC algorithm for each combination of monomers and their protonation states and/or conformers. The simulation box size, which defines spaces where the molecules are randomly placed in the initial step, was selected proportionally to each cluster. We use the strategy of a large amount of random

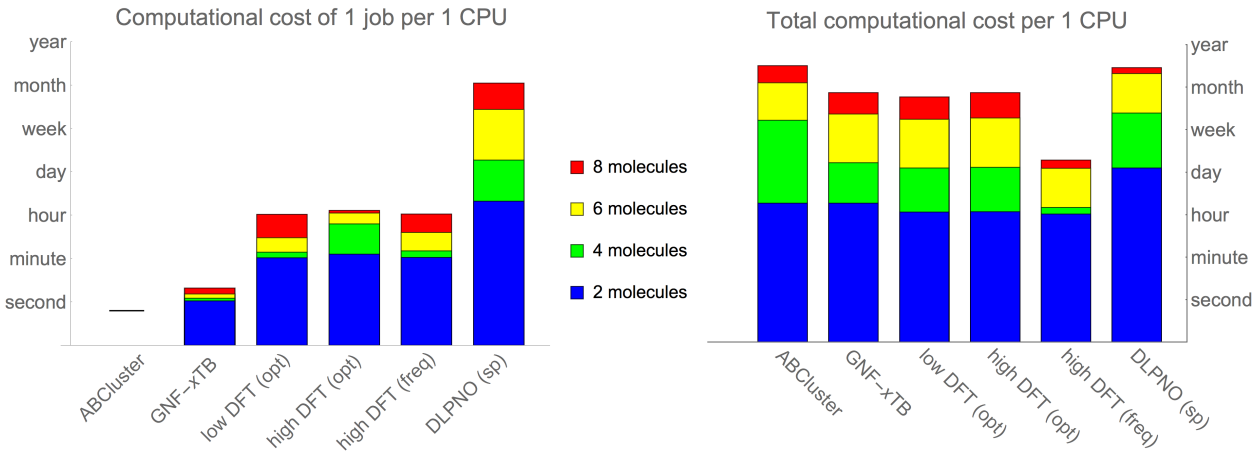


Figure 9: Computational costs of different configurational sampling steps for the sampling protocol used in this work. Different colors represent costs for different cluster sizes (see the legend in the middle). We also distinguish the computational cost of 1 job (left figure) and the computational cost multiplied by amount of jobs (right figure). Note the non-linear scale of the y-axis.

initial guesses (trial solutions) $SN = 3000$ and a small amount of generations $g_{\text{lim}} = 200$. (g_{lim} should not be lower than 100, otherwise the system will not have time to converge properly). One could also use the opposite approach as suggested by the authors of ABCluster: $SN \approx$ 20 - 100 and $g_{\text{lim}} \approx 100000$, which works better for atomic clusters or clusters containing small molecules.[22,23] We save large amounts of local minima $LM = 10^4$ for each combination of monomers, which are then treated at higher levels in the subsequent steps. The amount of scout bees (scout limit, *i.e.* the maximum number of generations that one minimum can last until it is replaced by a new random configuration) was set to $SC_{\text{bee}} = 4$. As shown in figure 9, the total computational cost of ABCluster run with the above mentioned parameters might be up to 1 cpu-year. However, for each combination of monomers, a separate run can be performed on different CPU. Moreover, since ABCluster is well parallelized, we can perform a calculations using more processors/cores. Therefore, a proper exploration of a PES for cluster containing eight molecules can be performed within 1 day provided that a sufficient number of processors/cores are available.

Next, we perform a GFN-xTB optimization for all structures saved from the ABCluster. We use the very tight optimization criteria. The optimization converges many different

input structures to the same minimum, which helps sort out redundant configurations (as described in the section **2.2.2 - Uniqueness, Filtering and Sampling**). Moreover, as shown in figure 9, the computational cost of this step is not critical, since performing a lot of short single jobs can be done in parallel. On the other hand, using this semi-empirical method might cause some relevant minima to be lost because the GFN-xTB PES slightly differs from the DFT PES.

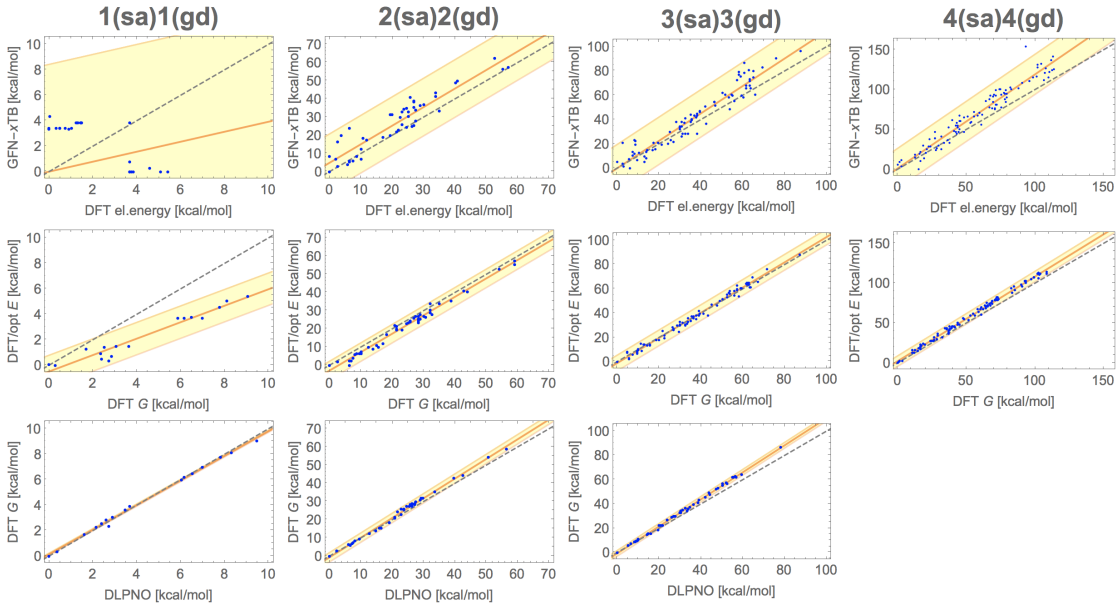


Figure 10: Energy correlations between different procedure steps (GFN-xTB energy → DFT [ωB97X-D/6-31++G(d,p)] electronic energy → DFT [ωB97X-D/6-31++G(d,p)] Gibbs free energy → Gibbs free energy at DLPNO//ωB97X-D/6-31++G(d,p) level) for four different molecular systems (see labels on top). Fit function (orange) and 95%-probability distribution area (yellow) are shown. (See Supporting Information for more details.)

For all subsequent steps, we performed statistical analysis of the correlations between the energies computed for all the minima found as part of our global minimum search. Figure 10 shows the correlations between the steps shown in table 1 (excluding the ABCcluster step, and the DFT pre-optimization step). The numerical results of this statistical analysis are shown in the Supporting Information. The first row of figure 10 shows that the relative energies of GFN-xTB correlate with the electronic energies of DFT optimized structures within an uncertainty $\approx 5N$ kcal/mol (see supporting information). Therefore, we filter out all

structures with a relative energy higher than $5N$ kcal/mol after the GFN-xTB optimization step.

Even though we delete redundant structures and filter out high-energy configurations, a prohibitively large number of structures still remains after the GFN-xTB step. Therefore, we perform a uniform selection/sampling of $\approx 100N$ structures for further DFT calculations from the whole filtered set of structures (see section **2.2.2 - Uniqueness, Filtering and Sampling**). Based on our experience, it is not yet clear, if it is better to sample a larger number of structures, or sample a smaller number of structures followed by another re-sampling based on the DFT results of the first sampled set.

The structures selected by the sampling are then analysed using DFT. To reduce computational cost, we recommend performing a DFT pre-optimization step at a lower level. In our case, we use the same functional as for the "high-level DFT" (ωB97X-D), but with a smaller basis set (6-31+G(d)), and loose optimization criteria. As shown in figure 9 and table 1, performing DFT pre-optimizations allows us to optimize an order of magnitude more structures compared to using just the high-level DFT method, at minimal additional cost. Moreover, the DFT pre-optimization also somewhat reduces the computational cost of the subsequent higher-level optimizations, as fewer optimization steps are needed.

After the pre-optimization, we optimize all structures with a relative energy lower than $2N$ kcal/mol using ωB97X-D/6-31++G(d,p) with very tight optimization criteria. Optimization and vibrational frequency analysis could be performed together, but we prefer to first optimize the structures, address problematic jobs (*e.g.*, those that did not converge, reached saddle points instead of minima), filter high-energy structures, and then perform vibrational frequency calculations for the remaining structures. We filter out high-energy structures using a threshold $1.7N$ kcal/mol (see Supporting Information). The frequency analysis also provides the possibility to check for imaginary frequencies, and thus see which structures need more optimization steps. The Gibbs free energy is then calculated using a quasi-harmonic approximation for vibrations with a frequency threshold of 100 cm^{-1}.

GUANIDINE	4	0.10	0.97	0.00	1.26	0.15
	3	0.00	0.05	0.00	0.42	0.00
	2	0.00	0.00	0.00	0.00	0.03
	1	0.00	0.00	0.02	0.01	0.00
	0	0.00	0.00	0.00	0.00	0.00
in [kcal/mol]		0	1	2	3	4
		SULPHURIC ACID				

Figure 11: Difference of DLPNO//ωB97X-D/6-31++G(d,p) Gibbs free energies (at 298.15 K) between the lowest minima found by the universal protocol and global minima shown in figure 6. The red color background highlights the minimum energy difference from global minimum greater than 0.5 kcal/mol.

Finally, on top of the high-level DFT structures with the lowest Gibbs free energies, we calculate the electronic energy correction using DLPNO. Figure 10 shows that the DLPNO//ωB97X-D/6-31++G(d,p) Gibbs free energy strongly correlates with the Gibbs free energy calculated solely using DFT. Therefore, just the few (2-3) lowest-energy clusters need to be treated with the memory-demanding and computationally expensive DLPNO method (see figure 9 and supporting information).

To summarize this section, we have developed the JKCS program and showed that it is able to find global minima compared to previous study (figure 7). We analysed the configurational sampling steps and stated parameters (for filtering *etc.*) to create a protocol dependent just on a cluster size (table 1). The protocol was applied for configurational sampling of sulfuric acid–guanidine system (figure 11). Figure 11 shows the energy difference between the lowest minimum found by our protocol, and the global minimum shown in figure 6. The figure shows that the protocol usually finds either the global minimum, or at least a structure energetically very close to it. However, there are also two structures, where the protocol fails by more than 0.5 kcal/mol. Nevertheless, this protocol represent a cost-effective approach for configurational sampling with an uncertainty of only a few kcal/mol. Especially for the larger clusters studied here, the other error sources of the computed quantum chemical

free energies can easily be several kcal/mol.

The protocol presented in this section is universal in the sense that all steps are simply functions of the number of molecules. Thus, it is easy to understand the complexity/system-size dependence. Users can adjust and apply this protocol to other systems as well. However, this might require considerable extra effort for systems very different from the one studied here, and users may thus have to adjust all configurational sampling parameters. The following key criteria should be checked when the protocol is applied to a new system: the amount/dimensionality of bonding patterns, the energy gain by addition of a new molecule, and the uniqueness criteria. These can be approximately compared to the sulfuric acid–guanidine system, and appropriate parameters for configurational sampling should then be rescaled when necessary.

3.3 - Symmetry Contribution

We utilized the program SYMMOL[60] to analyse the symmetry point groups of all global minima. The program SYMMOL tries to find the largest symmetry point group of molecular clusters with a pre-defined symmetry matrix deviation threshold. In this article, we varied SYMMOL tolerance thresholds between values 0.01 and 1.5 Ångström, and also used the options of either accounting or not accounting for the presence of hydrogen atoms in the symmetry matrix calculations. Unfortunately, the suitable symmetry matrix threshold varies with the cluster in question, and thus we assign point groups to the clusters based on a combination of SYMMOL results and our chemical intuition.

Table 2 shows clusters (excluding the monomer of sulfuric acid, σ(sa)= 2) which have global minima (as shown in figure 6) with a symmetry point group different from C_1. As shown in the table, higher symmetry leads to a higher Gibbs free energy. This is correct, because the symmetry number is a correction for over-counting all micro-states reachable by rotations of the cluster. In our case, all clusters have a rotational symmetry number between 1 and 3, except the very symmetrical $(sa)_4(gd)_4$.

Table 2: Point groups, rotational symmetry numbers and corrections to formation Gibbs free energy of symmetrical global minima from figure 6.

cluster	point group	rot. symm. number	form. free energy [kcal/mol]
$(sa)_2$	C_i	1	-
$(gd)_2$	C_{2v}	2	$-0.96 \rightarrow -0.61$
$(gd)_4$	S_4	2	$-8.51 \rightarrow -8.16$
$(sa)_2(gd)_2$	C_{2v}	2	$-65.05 \rightarrow -64.70$
$(sa)_4(gd)_1$	C_{3v}	3	$-56.49 \rightarrow -55.93$
$(sa)_4(gd)_3$	C_i	1	-
$(sa)_4(gd)_4$	T_D	12	$-165.04 \rightarrow -163.79$

Even though assigning a point group different from C_1 might cause another local minimum to become lower in free energy (thus becoming the new global minimum), the energy difference caused by variations in the rotational symmetry number is smaller than the uncertainty caused by approximations of anharmonic vibrations. Therefore, the search for global minima can still be carried out, and formation free energies of cluster can be calculated, without needing to make assumptions about symmetry.

4 - Conclusion

Configuration sampling of molecular clusters is important, for example, in atmospheric cluster distribution studies, because a key variable, the cluster evaporation rate, is exponentially dependent on free energies. Thus, energy differences greater than 1 kcal/mol cause differences in the order of magnitude of the evaporation rate.

We present a systematic method for performing configurational sampling of atmospherically relevant hydrogen-bonded molecular clusters. We develop and validate our sampling protocol based on clusters containing sulfuric acid and guanidine molecules, which have a large number of different bonding patterns.

Proton transfers reactions, which play a key role in the atmospheric chemistry, significantly complicate the process of searching for global minima, as accurately describing proton transfer requires a quantum chemical treatment (*i.e.*, can not be done using simple molecular

mechanics based on force field methods). In this article, we introduce a method for treating this problem by carrying out the molecular mechanics–based potential energy surface sampling using rigid molecules or ions corresponding to all possible combinations of acid/base protonation states (as well as different conformers of the monomers).

Further, we propose a sequence of pre-optimizations, re-optimizations, filtering, sampling/selection *etc.* processes allowing the generation of representative low-energy minima of molecular clusters at a minimal computational cost. We illustrate the application of this protocol to sulfuric acid–guanidine clusters of different sizes, and show that it is able to find global minimum structures within an uncertainity of around 2 kcal/mol. We also present all global minima which we have found during our research, and compare them to those presented in a previous study by Myllys *et al.*[3] By changing the parameters of the protocol, we can improve our configurational sampling, and thus increase the probability of finding the global minimum. A very important point and advantage of the configurational sampling approach presented in this article is that it does not require any information about minima of cluster size N in order to find the global minimum of cluster size $N + 1$.

Understanding the configurational sampling problem in greater detail allows us to study multi-component atmospheric clusters much more systematically and up to larger sizes than previously.

Acknowledgement

This research is supported by the European Research Council project 692891-DAMOCLES, Academy of Finland and University of Helsinki, Faculty of Science ATMATH project. We thank the CSC – Finnish IT Centre for access to computer clusters and also computer capacity from the Finnish Grid and Cloud Infrastructure (persistent identifier urn:nbn:fi:research-infras-2016072533). N.Myllys thanks the Jenny and Antti Wihuri foundation for financial support. J.Kubečka also gratefully acknowledges discussions with Filip Uhlík concerning

deep understanding of treatment of multi-local minima systems.

Supporting Information Available

- Supporting Information

  S1 Statistical Analysis of Configurational Sampling Steps

  S1.1 From GFN-xTB to DFT

  S1.2 From Electronic Energy to Thermochemistry at DFT Level

  S1.3 From Thermochemistry at DFT Level to Final DLPNO Corrected Results

  S2 Global Free Energy Minimum Structures

  S3 Gibbs free energy assuming all local minima

- global_minima.zip

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Kulmala, M.; Riipinen, I.; Sipilä, M.; Manninen, H. E.; Petäjä, T.; Junninen, H.; Maso, M. D.; Mordas, G.; Mirme, A.; Vana, M. et al. Toward direct measurement of atmospheric nucleation. *Science* **2007**, *318*, 89–92.

(2) Partanen, L.; Vehkamäki, H.; Hansen, K.; Elm, J.; Henschel, H.; Kurtén, T.; Halonen, R.; Zapadinsky, E. Effect of conformers on free energies of atmospheric complexes. *J. Phys. Chem. A* **2016**, *120*, 8613–8624.

(3) Myllys, N.; Ponkkonen, T.; Passananti, M.; Elm, J.; Vehkamäki, H.; Olenius, T. Guanidine: A highly efficient stabilizer in atmospheric new-particle formation. *J. Phys. Chem. A* **2018**, *122*, 4717–4729.

(4) Howard, A. E.; Kollman, P. A. An analysis of current methodologies for conformational searching of complex molecules. *J. Med. Chem.* **1988**, *31*, 1669–1675.

(5) Villamagna, F.; Whitehead, M. A. Comparison of complete conformational searching and the energy-optimized tree branch method in molecular mechanics calculations. *J. Chem. Soc.* **1994**, *90*, 47–54.

(6) Wales, D. J.; Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **1997**, *101*, 5111–5116.

(7) Kästner, J. Umbrella sampling. *WIREs Comput. Mol. Sci.* **2011**, *1*, 932–942.

(8) Chen, W.; Ferguson., A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comp. Chem.* **2018**, *39*, 2079–2102.

(9) Galvelis, R.; Sugita, Y. Neural network and nearest neighbor algorithms for enhancing sampling of molecular dynamics. *J. Chem. Theory Comput.* **2017**, *13*, 2489–2500.

(10) Deaven, D. M.; Ho, K. M. Molecular geometry optimization with a genetic algorithm. *Phys. Rev. Lett.* **1995**, *75*, 288–291.

(11) Hartke, B. Application of evolutionary algorithms to global cluster geometry optimization. *In Applications of Evolutionary Computation in Chemistry; Johnston, R. L., Ed.; Springer: Berlin* **2004**, *110*, 33–53.

(12) Judson, R. Genetic algorithms and their use in chemistry. *Rev. Comp. Chem.* **1997**, *10*, 1–74.

(13) Morris, G. A.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(14) Koskowski, F.; Hartke, B. Towards protein folding with evolutionary techniques. *J. Comp. Chem.* **2005**, *26*, 1169–1179.

(15) Jensen, F. *Introduction to computational chemistry*; John Wiley & Sons, Inc.: USA, 2006.

(16) Smellie, A.; Stanton, R.; Henne, R.; Teig, S. Conformational analysis by intersection: CONAN. *J. Comput. Chem.* **2003**, *24*, 10–20.

(17) Dieterich, J. M.; Hartke, B. OGOLEM: Global cluster structure optimisation for arbitrary mixtures of flexible molecules. A multiscaling, object-oriented approach. *Molecular Physics* **2010**, *108*, 279–291.

(18) Kanters, R. P. F.; Donald, K. J. CLUSTER: Searching for unique low energy minima of structures using a novel implementation of a genetic algorithm. *J. Chem. Theory Comput.* **2014**, *10*, 5729–5737.

(19) Temelso, B.; Morrison, E. F.; Speer, D. L.; Cao, B. C.; Appiah-Padi, N.; Kim, G.; Shields, G. C. Effect of mixing ammonia and alkylamines on sulfate aerosol formation. *J. Phys. Chem. A* **2018**, *122*, 1612–1622.

(20) Kildgaard, J. V.; Mikkelsen, K. V.; Bilde, M.; Elm, J. Hydration of atmospheric molecular clusters: A new method for systematic configurational sampling. *J. Phys. Chem. A* **2018**, *122*, 5026–5036.

(21) Karaboga, D.; Basturk, B. On the performance of artificial bee colony (ABC) algorithm. *Appl. Soft Comput.* **2008**, *8*, 687–697.

(22) Zhang, J.; Dolg, M. ABCluster: the artificial bee colony algorithm for cluster global optimization. *Phys. Chem. Chem. Phys.* **2015**, *17*, 24173–24181.

(23) Zhang, J.; Dolg, M. Global optimization of rigid molecular clusters by the artificial bee colony algorithm. *Phys. Chem. Chem. Phys.* **2016**, *18*, 3003–3010.

(24) Malloum, A.; Fifen, J. J.; Conradie, J. Structures and spectroscopy of the ammonia eicosamer, $(NH_3)_{n=20}$. *J. Chem. Phys.* **2018**, *149*, 024304–024304.

(25) Kumar, M.; Li, H.; Zhang, X.; Zeng, X. C.; Francisco, J. S. Nitric acid–amine chemistry in the gas phase and at the air–water interface. *J. Am. Chem. Soc.* **2018**, *140*, 6456–6466.

(26) Ma, X.; Sun, Y.; Huang, Z.; Zhang, Q.; Wang, W. A density functional theory study of the molecular interactions between a series of amides and sulfuric acid. *Chemosphere* **2019**, *214*, 781–790.

(27) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I. et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

(28) Yu, W.; He, X.; Vanommeslaeghe, K.; MacKerell, A. D., Jr. Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J. Comput. Chem.* **2012**, *33*, 2451–2468.

(29) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. Direct MP2 gradient method. *Chem. Phys. Lett.* **1990**, *166*, 275–280.

(30) Frisch, M. J.; Head-Gordon, M.; Pople, J. A. Semi-direct algorithms for the MP2 energy and gradient. *Chem. Phys. Lett.* **1990**, *166*, 281–289.

(31) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.* **1988**, *153*, 503–506.

(32) Saebø, S.; Almlöf, J. Avoiding the integral storage bottleneck in LCAO calculations of electron correlation. *Chem. Phys. Lett.* **1989**, *154*, 83–89.

(33) Head-Gordon, M.; Head-Gordon, T. Analytic MP2 frequencies without fifth order storage: Theory and application to bifurcated hydrogen bonds in the water hexamer. *Chem. Phys. Lett.* **1994**, *220*, 122–128.

(34) Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold, F. NBO Version 3.1.

(35) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(36) Knight, J. L.; Brooks, C. L. Validating CHARMM parameters and exploring charge distribution rules in structure-based drug design. *J. Comp. Theory Comput.* **2009**, *5*, 1680–1691.

(37) Grimme, S.; Bannwarth, C.; Shuskov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (Z = 1-86). *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

(38) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(39) Grimme, S. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. **2019**,

(40) Stewart, J. J. P. Optimization of parameters for semiempirical methods. V. Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *120*, 1173–1213.

(41) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **2013**, *120*, 1–32.

(42) Temelso, B.; Mabey, J. M.; Kubota, T.; Appiah-Padi, N.; Shields, G. C. ArbAlign: A tool for optimal alignment of arbitrarily ordered isomers using the Kuhn-Munkres algorithm. *J. Chem. Inf. Model.* **2017**, *57*, 1045–1054.

(43) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.* **1976**, *A32*, 922–923.

(44) Walker, M. W.; Shao, L.; Volz, R. A. Estimating 3-D location parameters using dual number quaternions. *CVGIP: Image Understanding* **1991**, *54*, 358–367.

(45) Stepto, R.; Chang, T.; Kraatochvíl, P.; Hess, M.; Horie, K.; Sato, T.; Vohlídal, J. Definitions of terms relating to individual macromolecules, macromolecular assemblies, polymer solutions, and amorphous bulk polymers (IUPAC Recommendations 2014). *Pure Appl. Chem.* **2015**, *87*, 72–120.

(46) Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **1901**, *6*, 559–572.

(47) Hotteling, H. Relations between two sets of variates. *Biometrika* **1936**, *28*, 321–377.

(48) Chaudhuri, B. B. How to choose a representative subset from a set of data in multi-dimensional space. *Pattern Recognit. Lett.* **1994**, *15*, 893–899.

(49) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(50) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B Condens. Matter.* **1992**, *46*, 6671–6689.

(51) Zhao, Y.; Truhlar, D. G. Design of density functionals that are broadly accurate for thermochemistry, thermochemical kinetics, and nonbonded interactions. *J. Phys. Chem. A* **2005**, *109*, 5656–5667.

(52) Myllys, N.; Elm, J.; Kurtén, T. Density functional theory basis set convergence of sulfuric acid-containing molecular clusters. *Comp. Theor. Chem.* **2016**, *1098*, 1–12.

(53) Elm, J.; Bilde, M.; Mikkelsen, K. V. Assessment of density functional theory in predicting structures and free energies of reaction of atmospheric prenucleation clusters. *J. Chem. Theory Comput.* **2012**, *8*, 2071–2077.

(54) Elm, J.; Mikkelsen, K. V. Computational approaches for efficiently modelling of small atmospheric clusters. *Chem. Phys. Lett.* **2014**, *615*, 26–29.

(55) Funes-Ardois, I.; Paton, R. GoodVibes: GoodVibes v1.0.1. **2016**, DOI: http://dx.doi.org/10.5281/zenodo.60811.

(56) Chon, N. L.; Lee, S.-H.; Lin, H. A theoretical study of temperature dependence of cluster formation from sulfuric acid and ammonia. *Chem. Phys.* **2014**, *433*, 60–66.

(57) Grimme, S. Supramolecular binding thermodynamics by dispersion-corrected density functional theory. *Chem. Eur. J.* **2012**, *18*, 9955–9964.

(58) Slanina, Z. *Contemporary theory of chemical isomerism (understanding chemical reactivity)*; Reidel Publ. Co.: Dordrecht, 1986.

(59) Goldman, M. J.; Ono, S.; Green, W. H. Correct symmetry treatment for X + X reactions

prevents large errors in predicted isotope enrichment. *J. Phys. Chem. A* **2019**, Articles ASAP.

(60) Pilati, T.; Forni, A. SYMMOL: a program to find the maximum symmetry group in an atom cluster, given a prefixed tolerance. *J. Appl. Cryst.* **1998**, *31*, 503–504.

(61) Gilson, M. K.; Irikura, K. K. Symmetry numbers for rigid, flexible, and fluxional molecules: Theory and applications. *J. Phys. Chem. B* **2010**, *114*, 16304–16317.

(62) Lloyd-Evans, D. J. R. Statistics and the symmetry groups of non-rigid molecules. *Mol. Phys.* **1966**, *10*, 377–380.

(63) Myllys, N.; Elm, J.; Halonen, R.; Kurtén, T.; Vehkamäki, H. Coupled cluster evaluation of the stability of atmospheric acid–base clusters with up to 10 molecules. *J. Phys. Chem. A* **2016**, *120*, 621–630.

(64) Riplinger, C.; Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013**, *138*, 034106.

(65) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J. Chem. Phys.* **2013**, *139*, 134101.

(66) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse maps – A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J. Chem. Phys.* **2016**, *144*, 024109.

(67) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(68) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron affinities of the first-row

atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(69) Liakos, D. G.; Sparta, M.; Kesharwani, M. K.; Martin, J. M. L.; Neese, F. Exploring the accuracy limits of local pair natural orbital coupled-cluster theory. *J. Chem. Theory Comput.* **2015**, *11*, 1525–1539.

(70) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H. et al. Gaussian 16 Revision A.03. 2016; Gaussian Inc. Wallingford CT.

(71) Neese, F. The ORCA program system. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73–78.

(72) Elm, J.; Bilde, M.; Mikkelsen, K. V. Influence of nucleation precursors on the reaction kinetics of methanol with the OH radical. *J. Phys. Chem. A* **2013**, *117*, 6695–6701.

Graphical TOC Entry